

REAL TIME ANALYSIS OF STREAMING DATA USING APACHE STORM



Group members:

Prashanth Kurella
Ojasvi Bhalerao
Abhinav Ralhan
Gaurav Malhotra
Anirudh Kumar Rao

Capstone Project

Documentation Report

16.04.2018

Group 34

NU Mentor: Dr. Ram Narayan Yadav, Mr. Ayan Nandy

NIIT Mentor: Mr. Manish Hurkat

Overview

The immense increase in big data applications and analytics has led to creation of specific softwares such as Apache Storm. It is also one of the primary reasons we chose to work on this project, to learn the concepts behind a big data architecture and learn how it works in application.

Here, we implement a live analysis on data and visualizing it. Our aim is to use real time Twitter streaming data for analysis and visualization of big data. Twitter data can be used for various purposes and since it is freely available and also in NoSql format, we chose to work with it. We aim to utilize this data in multiple ways as discussed further. We also create visualizations for all of them. The primary aim, however, is to understand the usage of real time data using Apache Storm and related technologies used in the process.

Goals

1. Sentiment Analysis of Tweets
2. Understanding the Trending topics
3. Geospatial Analysis of data
4. Realizing the reach of tweet

Technologies Used

I. Apache Storm



Apache Storm is an open-source distributed real-time computational system for processing data streams. Similar to what Hadoop does for batch processing, Apache Storm does for unbounded streams of data in a reliable manner.

- Able to process over a million jobs in fraction of a second on a node
- Integrated with Hadoop to harness higher throughputs
- Easy to implement and can be integrated with various programming languages.

II. MongoDB

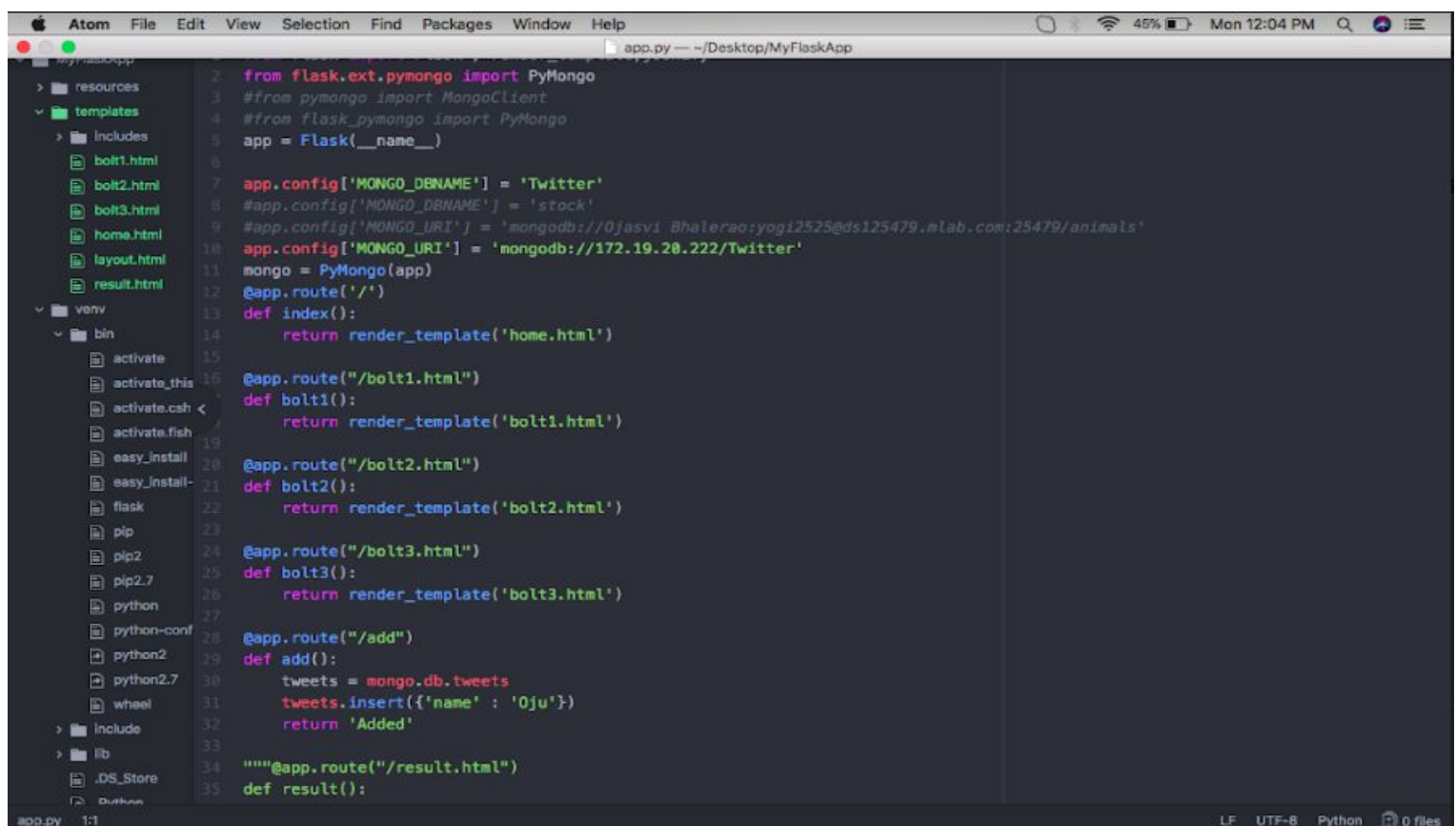
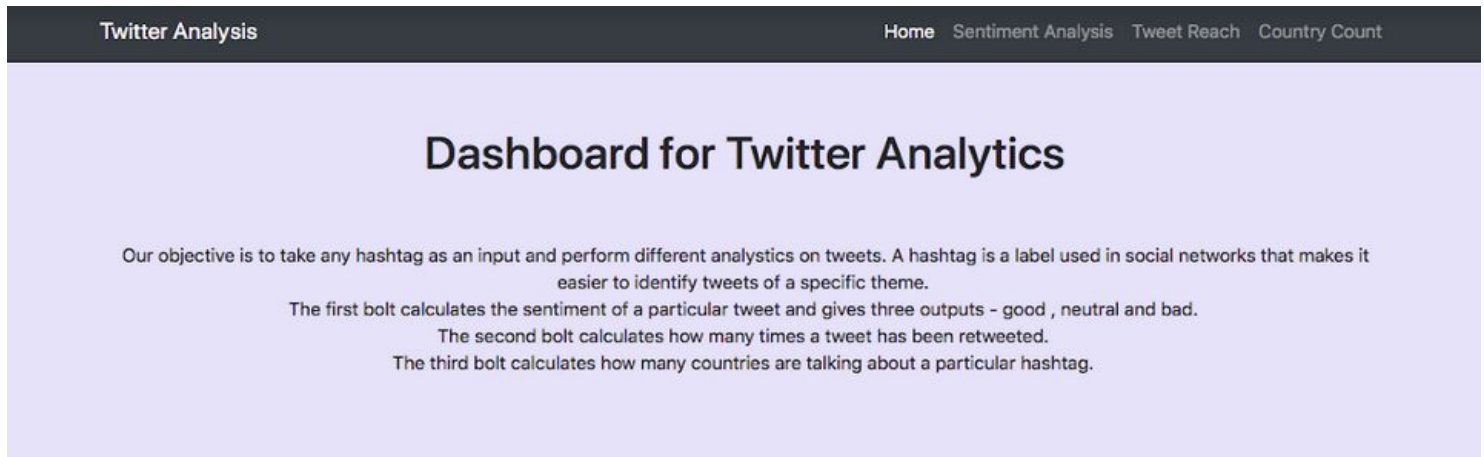


MongoDB is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas

- MongoDB has official drivers for major programming languages and development environments.
- MongoDB supports various features such as Indexing, Replication, Sharding, Aggregation, Load Balancing
- The language drivers are available under the Apache License, and it is available at no cost.

Progress

UI to display data



Sentiment Analysis

Sentiment analysis, also known as opinion mining is the process of determining the emotional tone behind a series of words, used to gain an understanding of the the attitudes, opinions and emotions expressed within an online mention.

#	Bad	Neutral	Good
---	-----	---------	------

Tweets Retrieved

```
prashanth@Flash-Ub:~/Desktop/apache-storm-1.2.1/bin$ ./storm jar /home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar twst.TwitterTopology
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/prashanth/Desktop/apache-storm-1.2.1/lib/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Running: java -client -Ddaemon.name= -Dstorm.options= -Dstorm.home=/home/prashanth/Desktop/apache-storm-1.2.1 -Dstorm.log.dir=/home/prashanth/Desktop/apache-storm-1.2.1/logs -Djava.library.path=/usr/loc
al/lib/opt/local/lib:/usr/lib -Dstorm.conf.file= -cp /home/prashanth/Desktop/apache-storm-1.2.1/*:/home/prashanth/Desktop/apache-storm-1.2.1/lib/*:/home/prashanth/Desktop/apache-storm-1.2.1/extlib/*:/h
ome/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar:/home/prashanth/Desktop/apache-storm-1.2.1/conf:/home/prashanth/Desktop/apache-storm-1.2.1/bin -Ds
torm.jar=/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar -Dstorm.dependency.jars= -Dstorm.dependency.artifacts={} twst.TwitterTopology
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/prashanth/Desktop/apache-storm-1.2.1/lib/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
352 [main] INFO o.a.s.u.TupleUtils - Enabling tick tuple with interval [1]
453 [main] WARN o.a.s.u.Utils - STORM-VERSION new 1.2.1 old null
469 [main] INFO o.a.s.StormSubmitter - Generated ZooKeeper secret payload for MD5-digest: -8232709895840966527:-7843596252621896891
543 [main] INFO o.a.s.u.NimbusClient - Found leader nimbus : Flash-Ub:6627
585 [main] INFO o.a.s.s.a.AuthUtils - Got AutoCreds {}
588 [main] INFO o.a.s.u.NimbusClient - Found leader nimbus : Flash-Ub:6627
584 [main] INFO o.a.s.StormSubmitter - Uploading dependencies - jars...
584 [main] INFO o.a.s.StormSubmitter - Uploading dependencies - artifacts...
585 [main] INFO o.a.s.StormSubmitter - Dependency Blob keys - jars : [] / artifacts : []
596 [main] INFO o.a.s.StormSubmitter - Uploading topology jar /home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar to assigned location: /home/pras
hant/Desktop/apache-storm-1.2.1/storm-local/nimbus/inbox/stormjar-987b0c75-d7dc-4b60-a27e-2519da7ebaa9.jar
Start uploading file '/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar' to '/home/prashanth/Desktop/apache-storm-1.2.1/storm-local/nimbus/inbox/s
tormjar-987b0c75-d7dc-4b60-a27e-2519da7ebaa9.jar' (4553268 bytes)
[*****] 4553268 / 4553268
File '/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar' uploaded to '/home/prashanth/Desktop/apache-storm-1.2.1/storm-local/nimbus/inbox/stormjar
-987b0c75-d7dc-4b60-a27e-2519da7ebaa9.jar' (4553268 bytes)
730 [main] INFO o.a.s.StormSubmitter - Successfully uploaded topology jar to assigned location: /home/prashanth/Desktop/apache-storm-1.2.1/storm-local/nimbus/inbox/stormjar-987b0c75-d7dc-4b60-a27e-251
9da7ebaa9.jar
730 [main] INFO o.a.s.StormSubmitter - Submitting topology Tweeter in distributed mode with conf {"storm.zookeeper.topology.auth.scheme":"digest","storm.zookeeper.topology.auth.payload":"-823270989584
0966527:-7843596252621896891"}
730 [main] WARN o.a.s.u.Utils - STORM-VERSION new 1.2.1 old 1.2.1
1014 [main] INFO o.a.s.StormSubmitter - Finished submitting topology: Tweeter
prashanth@Flash-Ub:~/Desktop/apache-storm-1.2.1/bin$
```


Submission of Storm Topology

```
prashanth@Flash-Ub:~/Desktop/apache-storm-1.2.1/bin$ ./storm jar /home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar twst.TwitterTopology
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/prashanth/Desktop/apache-storm-1.2.1/lib/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Running: java -client -Ddaemon.name= -Dstorm.options= -Dstorm.home=/home/prashanth/Desktop/apache-storm-1.2.1 -Dstorm.log.dir=/home/prashanth/Desktop/apache-storm-1.2.1/logs -Djava.library.path=/usr/loc
al/lib/opt/local/lib:/usr/lib -Dstorm.conf.file= -cp /home/prashanth/Desktop/apache-storm-1.2.1/*:/home/prashanth/Desktop/apache-storm-1.2.1/lib/*:/home/prashanth/Desktop/apache-storm-1.2.1/extlib/*:/h
ome/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar:/home/prashanth/Desktop/apache-storm-1.2.1/conf:/home/prashanth/Desktop/apache-storm-1.2.1/bin -Ds
torm.jar=/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar -Dstorm.dependency.jars= -Dstorm.dependency.artifacts={} twst.TwitterTopology
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/prashanth/Desktop/apache-storm-1.2.1/lib/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
352 [main] INFO o.a.s.u.TupleUtils - Enabling tick tuple with interval [1]
453 [main] WARN o.a.s.u.Utils - STORM-VERSION new 1.2.1 old null
469 [main] INFO o.a.s.StormSubmitter - Generated ZooKeeper secret payload for MD5-digest: -0232709895840966527:-7843596252621896891
543 [main] INFO o.a.s.u.NimbusClient - Found leader nimbus : Flash-Ub:6627
565 [main] INFO o.a.s.a.AuthUtils - Got AutoCreds []
568 [main] INFO o.a.s.u.NimbusClient - Found leader nimbus : Flash-Ub:6627
584 [main] INFO o.a.s.StormSubmitter - Uploading dependencies - jars...
584 [main] INFO o.a.s.StormSubmitter - Uploading dependencies - artifacts...
585 [main] INFO o.a.s.StormSubmitter - Dependency Blob keys - jars : [] / artifacts : []
596 [main] INFO o.a.s.StormSubmitter - Uploading topology jar /home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar to assigned location: /home/pras
hant/Desktop/apache-storm-1.2.1/storm-local/nimbus/inbox/stormjar-987b0c75-d7dc-4b60-a27e-2519da7ebaa9.jar
start uploading file '/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar' to '/home/prashanth/Desktop/apache-storm-1.2.1/storm-local/nimbus/inbox/s
tormjar-987b0c75-d7dc-4b60-a27e-2519da7ebaa9.jar' (4553268 bytes)
[*****] 4553268 / 4553268
File '/home/prashanth/eclipse-workspace/TwitterWork/target/TwitterWork-0.0.1-SNAPSHOT-jar-with-dependencies.jar' uploaded to '/home/prashanth/Desktop/apache-storm-1.2.1/storm-local/nimbus/inbox/stormjar
-987b0c75-d7dc-4b60-a27e-2519da7ebaa9.jar' (4553268 bytes)
730 [main] INFO o.a.s.StormSubmitter - Successfully uploaded topology jar to assigned location: /home/prashanth/Desktop/apache-storm-1.2.1/storm-local/nimbus/inbox/stormjar-987b0c75-d7dc-4b60-a27e-251
9da7ebaa9.jar
730 [main] INFO o.a.s.StormSubmitter - Submitting topology Tweeter in distributed mode with conf {"storm.zookeeper.topology.auth.scheme":"digest","storm.zookeeper.topology.auth.payload":"-023270989584
0966527:-7843596252621896891"}
730 [main] WARN o.a.s.u.Utils - STORM-VERSION new 1.2.1 old 1.2.1
1014 [main] INFO o.a.s.StormSubmitter - Finished submitting topology: Tweeter
prashanth@Flash-Ub:~/Desktop/apache-storm-1.2.1/bin$
```

Future Results

1. Write the remaining bolts/spouts and complete our Storm topology.
2. Debug and ensure that data is perfectly generated as per our use cases.
3. To add visualizations for all of our remaining use cases as defined above and add them to the UI.